The Impact of More Transparent Interfaces on Behavior in Personalized Recommendation

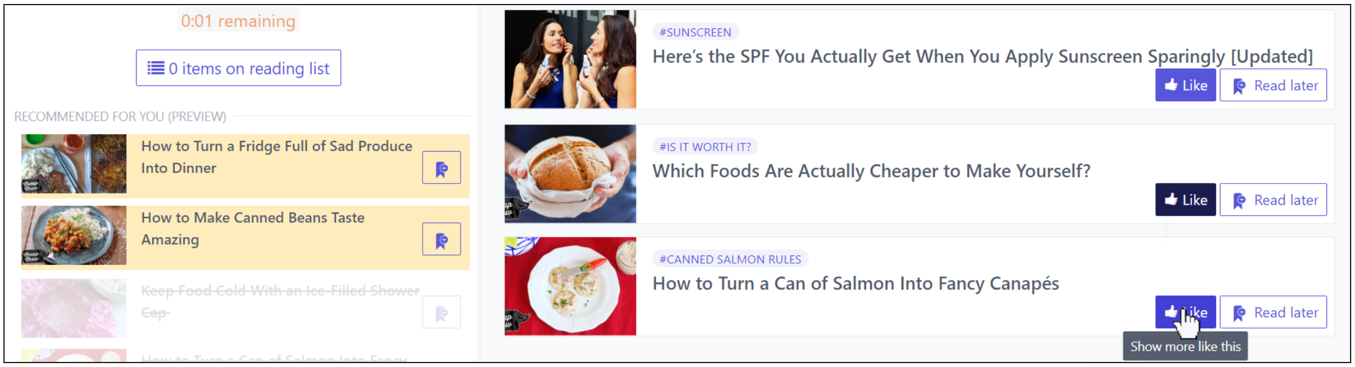
Tobias Schnabel[*]
Microsoft
Redmond, WA, USA
toschnab@microsoft.com

Saleema Amershi
Microsoft
Redmond, WA, USA
samershi@microsoft.com

Paul N. Bennett
Microsoft
Redmond, WA, USA
pauben@microsoft.com

Peter Bailey
Microsoft
Canberra, ACT, Australia
pbailey@microsoft.com

Thorsten Joachims
Cornell University
Ithaca, NY, USA
tj@cs.cornell.edu



How to Turn a Fridge Full of Sad Produce Into Dinner

How to Make Canned Beans Taste Amazing

~~Keep Food Cold With an Ice-Filled Shower Cap~~

~~How to Turn a Can of Salmon Into Fancy~~

Here's the SPF You Actually Get When You Apply Sunscreen Sparingly [Updated]

Which Foods Are Actually Cheaper to Make Yourself?

How to Turn a Can of Salmon Into Fancy Canapés

Figure 1: News system with interactive recommendations. In our experiment, we varied how and when people could see updates to the recommendations in response to their feedback actions on items. Here, people can preview the update that will be made to the recommendations by hovering over the "like" button before committing to actually clicking the button. The left panel also demonstrates the condition where *differences* to the person's previous recommendations are visible.

ABSTRACT

Many interactive online systems, such as social media platforms or news sites, provide personalized experiences through recommendations or news feed customization based on people's feedback and engagement on individual items (e.g., liking items). In this paper, we investigate how we can support a greater degree of user control in such systems by changing the way the system allows people to gauge the consequences of their feedback actions. To this end, we consider two important aspects of how the system responds to feedback actions: (i) immediacy, i.e., how quickly the system responds with an update, and (ii) visibility, i.e., whether or not changes will get highlighted. We used both an in-lab qualitative study and a large-scale crowd-sourced study to examine the impact of these factors on people's reported preferences and observed behavioral metrics. We demonstrate that UX design which enables people to preview the impact of their actions and highlights changes results in a higher reported transparency, an overall preference for this design, and a greater selectivity in which items are liked.

CCS CONCEPTS

• **Information systems** → **Personalization**; *Recommender systems*; • **Human-centered computing** → *User studies.*

KEYWORDS

personalization; machine learning; human-in-the-loop systems; control settings; user engagement

ACM Reference Format:
Tobias Schnabel, Saleema Amershi, Paul N. Bennett, Peter Bailey, and Thorsten Joachims. 2020. The Impact of More Transparent Interfaces on Behavior in Personalized Recommendation. In *43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3397271.3401117

[*] Work started while at Cornell University.

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401117

1 INTRODUCTION

state of the world

Personalized experiences are a core component of many online services, such as social media platforms, e-commerce websites, on-demand video services, and music providers. Recommendations and personalized news feeds have become an essential tool for driving user engagement, supporting choice making, and increasing feelings of trust [45]. Most of these personalized experiences are driven by Machine Learning (ML) and leverage implicit or explicit user feedback on individual items – such as viewing, liking, or saving items – in order to prioritize and recommend more relevant content in the future.

big but

One important challenge in building better personalization experiences is the question of how to implement user control in these systems, i.e., the ability for people to impact the recommendations they receive in an intended way. This is not only important for steering the algorithm away from unwanted recommendations, but it is also critical when little is known about a user in the beginning, or when long-term interests are insufficient for predicting preferences in new or other short-term contexts, e.g., when a person is trying to make a decision based on someone else's preferences. Moreover, user control is valuable because it has been shown to increase overall satisfaction with the system [17, 21], improve people's trust in the system [16, 25], and has also been associated with the intention to continue to use the system [32, 33].

Despite its importance, there has been relatively little research into how to improve user control in a dynamic personalization loop where people provide feedback on individual items. We seek to explore interfaces that will support people in answering the question: "*What impact will my actions have on my future recommendations?*". Two factors of the interface design that directly affect this question are (i) the timing or *immediacy* of the system response and (ii) the *visibility* of the system response. Immediacy – the decision when to present a response – is essential for understanding what the effects are because it provides the temporal link to tie actions with reactions. Similarly, the visibility of the response, i.e., how the system presents changes in the updated content is important because it is tied to the question of what changed in the recommendations. In the following, we consider a large spectrum of possible points in time to update for immediacy: showing a preview before a person chooses to follow through with an action (Preview), immediately after an action (Instantaneous), delayed after some number of actions (Delayed), and never (Never; i.e., showing a static set). We also vary visibility of the response by highlighting (Diff) or not highlighting (Refresh) what has changed.

Therefore we did…

In this paper, we present the results of both a in-lab and an online task-based study with an experimental news system offering personalized recommendations (Figure 1). Given its common adoption in current real-world recommender systems, we focus on "likes" as the primary feedback action by which users can update the recommendation panel. For this news recommender with personalized recommendations, people were given a fixed time-frame in which they had to compile a reading list of news articles they intended to read later (e.g. quickly choosing articles to download before boarding a plane and losing connectivity). For example, in Figure 1, on the right we see the user is hovering over "like" for the last article on salmon canapés. On the left, we see a preview of how recommendations will be updated if the user does click "like". If they click, the new recommendations highlighted in yellow will be added to the personalized recommendations for the user and those that are grayed out and in strikethrough font will be removed from the recommendations. This illustrates our Preview-Diff condition which previews the impact a "like" action will have on the recommendations with differences highlighted.

Our qualitative and quantitative study indicate the Preview condition is both preferred and increases selectivity in implicit behavioral metrics which may lead to more accurate personalization overall. Similarly, we find that the Diff condition is preferred and plays an important role in modulating attention to updates.

2 RELATED WORK

Our work intersects and connects work from various areas within interactive recommender systems and Machine Learning. We first provide a brief overview of work on interactive recommender systems, before turning to the two specific aspects of user control in recommender systems we study in this paper – the immediacy and visibility of system updates.

2.1 Interactive recommender systems

With a rich literature in this area, we focus on the most relevant systems and concepts in this section, but refer to current surveys [15, 18] for a more extensive review. Interactive recommender systems can roughly be grouped by whether or not they allow users to influence the underlying recommender model only temporarily or permanently. Examples for temporary operations are search, filters [36], or other slicing or zooming operations [47]. Methods in the second group that make permanent changes to the underlying recommender model typically employ a rich set of possible interaction patterns. On the richer end of the interaction spectrum are recommender systems where people give feedback via manual controls such as filters or sliders to specify the relative importance of certain attributes [5, 17, 41]. For example, in a system called TasteWeights [5], attributes are aggregated into concepts, and users can adjust weights on attributes (e.g., artist) as well as concepts (e.g., genre). A more indirect approach is to visualize importance as closeness to attributes, and let users specify weights implicitly via the 2-D position of attributes in a preference space [3, 30]. For example, the scientific article recommender by Bakolov et al. [3] places attributes on concentric circles colored by importance and provides users with editing capabilities, such as adding or removing them, or specifying semantic relationships between attributes. The core limitation of all attribute-based control methods is that they require that items and preferences can easily be described in terms of interpretable attributes (e.g., price, genre), which is hard for more complex items (e.g., images, natural language). Moreover, attribute-based feedback can be challenging to incorporate into existing ML-based algorithms since most supervised methods expect input in the form of item-label pairs, which is why many of the systems above rely on hand-crafted scoring functions.

Rather than focus on attribute-based control mechanisms (e.g. facets, topics, etc.) that are not highly utilized components of current recommender systems, we focus on "likes" as an instance of the larger class of item-based control methods. Item-based feedback has a long history in personalization and recommender systems [35]. The main design factor in item-based control is the question of granularity (e.g., binary or ordinal). Previous studies have shown that people typically prefer coarser rating scales because they are less effortful [42] and many large online platforms and content providers such as Facebook, YouTube or Netflix elicit mainly binary feedback. Moreover, having higher granularity than binary ratings does not necessarily produce more accurate recommendations [7]. For these reasons, we focus on binary feedback actions in this paper.

Regarding user experience, studies show that having user control in recommender systems makes people more satisfied with the suggestions that the system produces [17, 21]. Also, studies typically found increased engagement when controls were present as measured by the time people interacted with the personalized system [19, 31]. Interestingly, even the mere presence of feedback controls in a personalized system can already make people more satisfied with the content it presents [45], independent of whether the controls have an effect or not. Moreover, when people engage with feedback controls it has been shown to make them more likely to follow the system's recommendations [41]. Despite the prevalence of item-based feedback controls online, there are few studies that look at how people engage with such controls to customize the personalization. In one example, Eslami et al. [11] study how people reason about their feedback actions in Facebook's news feed and find that they engage in complex sense-making and feedback strategies.

2.2 Immediacy of system updates

The immediacy of updates to personalized experiences, i.e., the question of when to present users an update (a new prediction and/or model based on incorporating the feedback) after they perform a feedback action, has not received a lot of prior attention. We consider four settings in this paper – previewing, immediate, delayed and no updates. Traditionally, many real-world recommender services only update their personalized predictions with a significant delay [1]. This delay is often due to considerations arising from the ML-based algorithm, since many ML-based techniques perform what is known as *batch learning*, where the model is only inferred once from the data and then deployed. There are, however, systems that make real-time updates to their recommendations, such as the work by Wu et al. [50] or bandit-style algorithms [23]. In a smaller crowd-sourced study, Schaffer et al. [37] compare immediate and no updates for a movie recommender system. They find that in the immediate condition, people are more likely to provide additional ratings to the recommender system. However, they were explicitly asked to provide feedback to the recommender – a major difference to the task-based setting in our study where feedback was voluntary. Previewing is often used in the context of direct manipulation interfaces, such as bolding or applying a filter [43]. Related to the question of when to update the UI is the issue of response times (e.g., [28, 39]). Typically, people prefer shorter response times when interacting with an interface, and longer response times often lead to less engagement with it [14]. However, it is unclear to what extent this would translate to personalized experiences [40].

2.3 Visibility of system updates

As many personalized experiences update recommendations only on page refreshes or between sessions, the question of how to surface changes has not received much attention. One of the few interfaces that makes updates visible is CueFlik [12], an interface for image search, that surfaces updates after a feedback action via a reshuffling animation. Related to this is MrTaggy [19], an interface for exploratory search using votes on social tags as feedback actions. There, changes are both shown through a reshuffling animation as well as color-coded bars that indicate whether a search result was new. Another example is the interface of Elucidebug [22] for e-mail classification, where updates to the model were surfaced as up or down arrows.The system of Qvarfordt et al. [34] uses bar charts to show percentages of new and changed ranking results during document retrieval. Perhaps closest to the highlighting mechanism we employ is the system for Tweet recommendation by Waldner and Vassileva [48]. In it, the system highlights recommended items in bright yellow. However, this is done statically and not to indicate any updates to the recommendations.

In contrast to previous work, we study control and visibility in the context of common feedback actions ("likes") and focus on reflecting the system's response to the user's actions solely in *when* recommendations are updated and *how* they are displayed. Thus unlike the majority of the related work, we study how to introduce more control and transparency directly into common recommendation experiences rather than argue for complex UX redesigns of those experiences. Additionally we present the first study that investigates how people's behavior changes when they can *preview* the impact their next action would have on the recommendations they are receiving.

3 RESEARCH QUESTIONS

We explore two important factors of user control in interactive recommender systems pertaining to the way the system allows people to observe consequences of their actions, (1) the immediacy of a system's response via new recommendations that incorporate user feedback actions, i.e., when a system's response occurs; and (2) the visibility of the system's response to user feedback actions, i.e., how changes are surfaced to people. We examine four decreasing levels of immediacy, subsuming common updating schemes of ML-based recommender models as well as typical usability-based paradigms [26]: Preview, Instantaneous, Delayed, and Never. Preview enables people to see the impact of their feedback action before their action is carried out (which may change a person's decision to act). Instantaneous is when the system responds immediately after a feedback action has been performed. Delayed is when the system only updates after several actions. The Never setting simply means that the personalized content is held constant throughout a session. The latter three are most common in real-world recommender systems. For example, Amazon offers both instantaneous recommendations when clicking on an item as well as session-based recommendations that are updated after a few items have been browsed. Never is typical in streaming providers that would update recommendations independent of user visits, e.g., once a day.

Research Questions. People can decide where to put them, more commonly stated in Introduction or Method sections.

We also study two levels of visibility: one in which changes to the recommendations are explicitly displayed and highlighted, and one in which changes are not explicitly highlighted (but could be inferred by the user if they recall the previous set of recommendations). We ask three research questions covering the different aspects of user experience and system success. Our first one is regarding overall preferences

RQ1 What settings of immediacy and visibility do people prefer overall?

Closely related to this is the question of how this affects user-centric outcomes, such as feelings of control, transparency and user satisfaction with the personalized system.

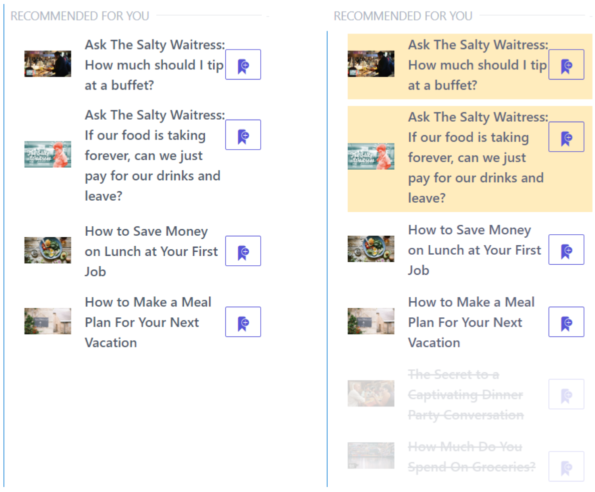
RQ2 How does the immediacy and visibility of a personalized system's responsiveness to user feedback actions affect
(a) feelings of control?
(b) perceived transparency?
(c) user satisfaction?

Regarding immediacy, our hypothesis is that people would perform best when immediacy is greatest. This is based on classic research on response time that demonstrates that longer response times negatively impact user satisfaction and productivity [39]. We also expect that this will carry over and positively impact feelings of control, perceived transparency, and user satisfaction. For visibility, we expect better visibility to also help people evaluate more quickly whether their action had the desired consequences. The importance of this evaluation step is highlighted in multiple interaction frameworks, such as Schön's reflection-in-action [38] and Norman's Seven stages of action [29]. Given this importance, we expect that higher visibility of what has changed to be preferred by people, with the potential to also improve perceived transparency and overall satisfaction.

Our last research question revolves around how much the different factors influence implicit user behavior. Tracking common metrics of user engagement, the quantities allow us to assess different facets of the overall experience. We therefore ask:

RQ3 How does the immediacy and visibility of a personalized system's responsiveness to user feedback actions affect
(a) engagement with the feedback controls?
(b) hovering vs. clicking on the feedback controls?
(c) engagement with the recommendations?
(d) overall engagement with the content?

These questions will be studied in tandem with the other research questions using the experimental news system described in the next section. To gain a more complete perspective, we conduct both a qualitative in-lab study as well as a quantitative online study whose results will be presented in Section 7. The qualitative in-lab study will allow us to interpret interaction signals from the quantitative study results in the right manner while the quantitative online study allows us to add statistical support to themes discovered in the in-lab study.



RECOMMENDED FOR YOU

Ask The Salty Waitress: How much should I tip at a buffet?

Ask The Salty Waitress: If our food is taking forever, can we just pay for our drinks and leave?

How to Save Money on Lunch at Your First Job

How to Make a Meal Plan For Your Next Vacation

(a) refresh

RECOMMENDED FOR YOU

Ask The Salty Waitress: How much should I tip at a buffet?

Ask The Salty Waitress: If our food is taking forever, can we just pay for our drinks and leave?

How to Save Money on Lunch at Your First Job

How to Make a Meal Plan For Your Next Vacation

~~The Secret to a Captivating Dinner Party Conversation~~

~~How Much Do You Spend On Groceries?~~

(b) diff

Figure 2: In the Refresh setting (a), differences were not visible, whereas in the Diff setting (b), new articles were highlighted and articles that would be removed were grayed out.

4 SYSTEM DESIGN

We conducted our studies with an experimental news system that allowed people to give feedback on items by liking them, similar to the functionality available in most popular news feeds. This news system was the same in both the in-lab study as well as the crowd-sourced study. The design of the news system was driven by two factors: (i) Ecological validity: the feedback actions available should be typical of those available in personalized services and the interaction flow should resemble prior experiences. (ii) Consistency: the interface should support all immediacy and visibility conditions without larger changes. For the design, we iterated repeatedly on prototypes with feedback from all authors as well as pilot users until all the desiderata above were met.

Methods: design goals

The basic design of the news system can be seen in Figure 1. The main panel supported basic browsing functionality such as paging, scrolling as well as clicking on articles to see the full article text and was the same across all conditions. People could also add articles to their reading list via the "Read later" button, and see and modify what was on their reading list. What was also common to all conditions was the descending countdown timer and the reading-list button on the top left. Only what was shown in the "Recommended for you" panel on the left was varied between the different conditions, and was visible all the time.

The system was populated with articles from Lifehacker[1], a news blog featuring articles about life hacks and productivity tips. In total, we had 684 articles crawled from the period between April 30, 2018 and July 16, 2018. We chose this content as it offered a large variety of possibly interesting topics to people, did not require prior expertise, and was almost exclusively evergreen content ensuring that articles wouldn't become obsolete in any subsequent study.

4.1 Implementation of settings

We mapped the different settings of visibility of an update, i.e., the system updating what was shown in the "Recommended for you" panel, in the following way:

[1] http://lifehacker.com

Refresh. In this setting, the list of recommendations simply *refreshes* entirely on an update. As Figure 2(a) shows, this also means that it is entirely up to the user to determine what has changed.

Diff. This setting makes the *differences* to the previous list of recommendations visible. New items are highlighted in yellow, and items that will be removed are grayed and crossed out. Unless the differences were displayed as part of a preview, they were slowly faded out after five seconds in order to make it clear that the changes were permanent. This setting is also shown in Figure 2(b).

The different immediacy settings determined when an update to the "Recommended for you panel" would happen.

Preview. Whenever people would mouse over the like button next to an article, a preview of how the recommendations would update after clicking like was displayed in the "Recommended for you" panel (cf. Figure 1). When ending the mousing over without a like click, the recommendation panel changed back to its prior state.

Instantaneous. Here, the recommendation panel was updated immediately after a click on the like button.

Delayed. The "Recommended for you panel" update was delayed until three new articles had received likes since the last update.

Never. Replicating the common experience in which recommendations are only updated after a session has been completed, the "Recommended for you" panel always showed the same four articles, and did not react to any feedback action. We picked the four most popular articles of the set, measured by the number of comments they received on Lifehacker as page views were not available.

Crossing immediacy and visibility yields a total of seven combined settings that we consider in this paper. Table 1 shows a summary of each of the settings. Note that in the Never setting of immediacy, visibility cannot be varied because no updates are made to the list of recommendations.

4.2 Personalization algorithm

The algorithm we used is a typical instance of content-based recommendation, namely nearest-neighbor in a learned embedding space, an approach whose robust performance has been validated empirically many times (e.g., [24, 46]). The algorithm was implemented in JavaScript and ran directly in the users' browsers, resulting in instantaneous updates. More specifically, the personalized recommendations were aggregated from the items that people liked via nearest-neighbor search in the embedding space which was created as follows. We extracted each article's full text, and mapped it to a 50-dimensional vector embedding via Latent Semantic Indexing [9]. Using an article's vector, we then retrieved the most similar articles to it in the embedding space as measured by cosine similarity. This was done for all articles that had been liked in a session. The final list of recommended articles was then compiled by filling up the four available slots in the recommendation panel in a round-robin fashion. Note that although collaborative filtering techniques are a popular choice for many industrial applications, they were not applicable to our scenario since we (i) did not have any interaction signals of other users with the news items and (ii) saw each user for the first (and last) time in our study. In our qualitative study, lab participants reported the personalization algorithm produced adequate and useful recommendations.

4.3 User Task

The main task for participants was to imagine they were at an airport and had five minutes to compile a reading list with interesting articles for their flight where they would have no internet. This was done in order to motivate the necessity of compiling such a list, rather than reading articles right away. The main prompt read: *"You have 5 minutes to find enough interesting articles for your flight and add them to your reading list."*

While an unconstrained time scenario in recommender systems usage may appear most natural, we attempt to strike a balance between natural and the need to hold as many things as possible constant to avoid having additional confounds. We chose to control for time for several reasons. First, for our online crowd study, it is recommended to set up a tasks in a way in online work markets that completion in good faith is as effortful as unfaithful completion [20]. Furthermore, there are many common tasks that are also time-constrained in nature, for example picking a movie to watch right now, booking flights whose availability is likely to change, or compiling a playlist with songs for offline usage.

After reading the task instructions, people were shown a brief 2 minute video that explained the basic functionality of the news system they were going to interact with next. Each video was adapted to only explain the functionality that was available in the current condition. The liking functionality was briefly explained, but received no further treatment in order to minimize priming effects. The participants then were taken to the page with the news system (cf. Figure 1) and were shown a brief message once the five minutes were up.

5 QUALITATIVE STUDY

We now turn to the details of the qualitative in-lab study. The goals of this qualitative study were to obtain rich feedback about user preferences and attitudes towards using the various interfaces as well as to understand how to interpret behavioral signals that we obtained in our large scale online study in the next section.

5.1 Participants

We recruited 11 participants via email lists containing employees of a large technology company. Participants were from various backgrounds spanning designers, engineers and administrators with a mean age of 42.8 years ($SD = 10.2$). Seven participants identified as female and four as male. Participants received a $25 Amazon gift card as compensation for the hour-long study.

5.2 Procedure

Before arriving at the lab, participants were asked to sign a consent form. Upon arriving, they completed a brief demographic survey and were then presented with the task instructions as outlined in Section 4.3. In addition, we asked them to think aloud as they were completing the task and as is common in usability testing. This helps capture the cognitive processes in parallel with the

Immediacy	Visibility	Description
Preview	Refresh	Before the like is pressed, preview the new list (but no diff).
	Diff	Before the like is pressed, preview the differences in what will change.
Instantaneous	Refresh	Right after the like is pressed, show the new list (but no diff).
	Diff	Right after the like is pressed, update the list and show what the differences are.
Delayed	Refresh	Every three likes, update and show the list (but no diff).
	Diff	Every three likes, update the list and show what the differences are.
Never	n/a	Never update the list

Table 1: All feasible combined settings of immediacy and visibility.

observed actions. For the study, we used a within-subject mixed factorial design with two blocks. Conditions within each block were randomized to mitigate learning effects. The first block consisted of three conditions where visibility was held constant after choosing it at random (Refresh or Diff). Within these three conditions, each participant interacted with Preview, Instantaneous and Delayed in random order. We excluded Never from this study because of its obvious lack of interactivity. After this, we changed the visibility of the last immediacy setting for the last condition and block. Each condition was followed by a brief semi-structured interview asking about their attitude towards the interface. Upon completing a block, we asked for preferences among the conditions they interacted with. In total, each participant repeated the user task four times with a different interface and a different set of articles.

In general, participants found the setup clear and realistic. Without being prompted, two participants commented on the user task: *"This is usually what I do at the airport."* (P16) and *"This scenario is very real to me. I have done this before."* (P9).

6 QUANTITATIVE STUDY

In addition to our qualitative in-lab study, we conducted a crowdsourced online study that studied behavior more quantitatively and with a larger set of participants. In this between-subjects study, each participant was randomly assigned one out of seven conditions.

6.1 Participants

We recruited 604 participants from Amazon Mechanical Turk. Participants had to be from a US-based location and had to have a previous task approval rate of 95%. Furthermore, we required that participants were using a recent browser that would render the interface correctly. The required content was cached locally in order to minimize the impact of varying internet speeds. We paid participants $1.60 for completing the experiment that lasted about 10 minutes on average, resulting in an approximate rate of $8 per hour. Participants had a mean age of 35 years (SD = 10, min = 19, max = 69), with 62% indicating male, 37% female, and 1% choosing not to answer.

6.2 Procedure

Each participant was randomly assigned one of the seven conditions of Table 1 at the start of the experiment. We instrumented the news system with common trackers of user engagement. We logged all mouse clicks and hovers on recommended items as well as all interactions with the like buttons and additions to the reading list. In the post-task survey, we asked for demographic information (age and gender) and usage of news sites. We also included an attention check question (*"Please select the option 'Strongly Agree' "*), common in Mechanical Turk questionnaires, to detect people that were only skimming through the questions.

7 ANALYSIS AND RESULTS

Here we discuss our key findings supported by qualitative and quantitative evidence from our lab and online studies.

7.1 Analysis

7.1.1 Qualitative study. Two authors analyzed participant comments from our think-aloud lab study by iteratively grouping comments to identify themes based on our research questions. The grouping process was repeated until consensus was reached. All themes and comments we present in this section are therefore representative of preferences and perceptions of multiple participants.

7.1.2 Quantitative study. In total, we had 601 participants complete the study, corresponding to about 85 participants for each of the seven conditions. From the 601 responses, we filtered out all responses where participants failed the attention check (six responses), as well as participants that showed manipulative behavior in the form of rapid clicking (34 responses). The latter was defined as having three or more consecutive actions of the same type (liking or readlisting) in less than 2.5 seconds. After the filtering step, we had 567 valid responses. We used ANOVA tests on continued-valued measures, and Kruskal–Wallis tests on responses that had an ordinal scale. We used Tukey's test for post-hoc analysis at $p = 0.05$.

7.2 Key Findings

Preview was strongly preferred for decreasing decision anxiety and increasing a sense of control over the system's behavior. Overall, participants in our lab study expressed a strong preference for the Preview condition in terms of the immediacy dimension (8 out of 11 participants). Several themes emerged as impacting participant preferences as evidenced by multiple participants commenting on similar attributes. First, the Preview condition, which allows users to preview the consequences of their actions before committing to any action, appeared to greatly reduce decision anxiety among our participants. Decision anxiety relates to concerns about actions (likes in this case) permanently steering

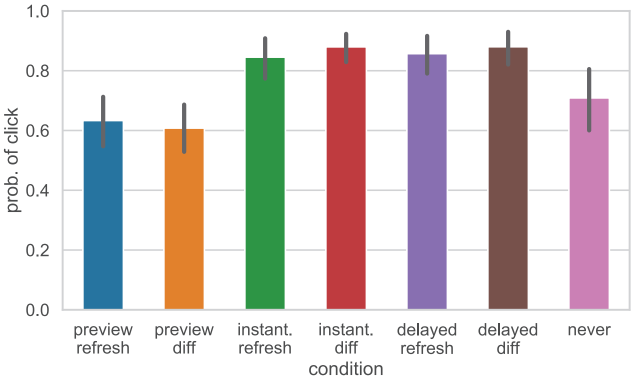


Figure 3: Probability of clicking on the like button after having hovered over it for more than one second. Data from crowd study (567 users).

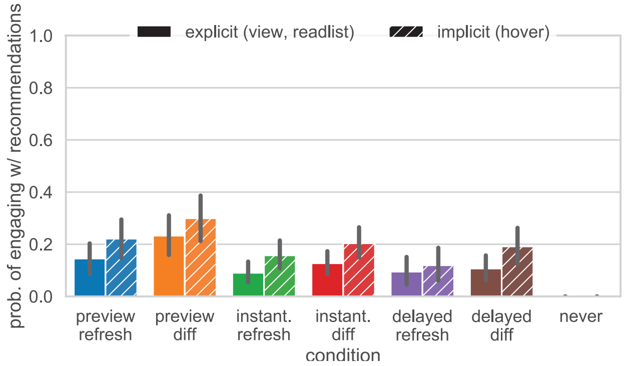


Figure 4: Probability of engaging with recommendations after an update to the recommendations (clicking on or read-listing a recommendation for explicit; or hovering for 1s+ for implicit) within 3 seconds after an update. Data from crowd study (567 users).

a system's recommendations in the wrong direction with no means for recourse. For example:

- *"Mentally, I don't have to make a decision. I hover over it and there is no commitment."* (P16)
- *"I didn't have to think too much – Is it gonna mess up my recommendations?"* (P33)
- *"With the [Preview] interface, I didn't necessarily have to commit to a click which felt nice."* (P7) This participant went on to say that their decision anxiety around clicking extended to the recommender systems they used in everyday life (e.g., *"I am reluctant to share my preferences. I do not want to 'freeze' in my preferences. If I find a hip hop song and I click like, it might assume I am a hip hop fan."*

Observations from our think-aloud lab study showed that this decreased decision anxiety often manifested as participants leveraging the preview through hover functionality to "peek" at recommendations and only committing via clicking when the recommendations were deemed to be of interest:

- *"I kinda get a quick peek if it is producing recommendations that are relevant."* (P35)
- *"I hovered over like to bring up possibly related articles. When they weren't interesting, I didn't click like."* (P31)
- *"It was kind of interesting to see in a non-committed way what would happen … I could kind of see – do I want to give that kind of signal or not?"* (P15)

This peeking behavior with the Preview condition was also evident in our online study where we found differences in participant selectivity in liking items across conditions. Figure 3 displays the probability that a hover event of one second or longer over a like button is followed by a click on that button. In general, most long hovers are indeed followed by a click as the probability is greater than 0.6. However, the likelihood of clicking like after hovering differs significantly across conditions (an ANOVA shows $p < 10^{-8}$). Post-hoc pairwise comparisons reveal that in the Preview conditions, participants are significantly less likely to follow through with a click than in the Delayed or Instantaneous conditions. In other words, participants demonstrated greater selectivity in which items they like in the Preview conditions.

Moreover, when participants did in fact commit to liking an item in the Preview condition, they were more likely to engage with the updated recommendations (e.g., clicking on or read-listing an item) than after updates in other conditions (see Figure 4). This increased engagement with the recommendations along with the peeking behavior we observed suggests that the Preview condition allowed participants to in effect "see into the future" and proceed only if they deemed an action as bringing value to their task. Participants in our lab study commented on this increased sense of control resulting from a better understanding of the effects of their actions (aligning closely with the established usability principle of keeping people informed through timely and appropriate feedback necessary to make appropriate future decisions [27]):

- *"[I] really like visibility into the effect of my actions."* (P31)
- *"I felt more informed what I was doing … Its nice to see how my feedback is being interpreted."* (P33)

The Preview condition was, however, not universally preferred. Of the three participants who preferred other conditions in our lab study (two preferred immediate while one preferred delayed), the main reasons they gave related to the cognitive costs of frequent updates, for example through being distracted or state keeping:

- *"It took a bit too much focus…"* (P15)
- *"[…] it was a little bit distracting."* (P35)
- *"I don't want to see every change in my recommended items… I don't have that much time or capacity to keep everything in memory."* (P2)

Delayed was least preferred and confounded traditional behavioral indicators of engagement. Although the Delayed condition is comparable to many real-world recommender systems where explicit user preferences are often not tied to immediate feedback or recommendation updates, the Delayed condition was least preferred by our lab study participants with 9 out of 11 choosing it as their least favorite condition (note that we only tested Preview, Instantaneous, and Delayed in our lab study, whereas our online study also included the Never condition for reference).

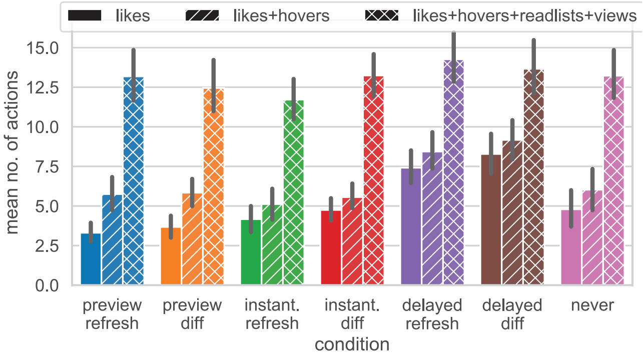


Figure 5: Average number of like actions (solid), like actions plus long hovers (1s+) on the like button (hatched), all item-based engagement (cross-hatched) across conditions. Data from crowd study (567 users).

Participant comments revealed this was often due to frustration at the lack of responsiveness and confusion about what was happening, particularly in comparison with the other conditions.

- *"A mile of frustration of nothing happening ... My brain expected change. Not only would it take a few interactions, I never knew how many interactions."* (P9)
- *"Less clear to me when I should click like on something ... I didn't get any immediate feedback."* (P33)
- When interacting with the Delayed condition for a participant whose random assignment gave the preview condition first: *"Now I really want the hover feature!"* (P40)

Surprisingly, this frustration at the lack of responsiveness often manifested as participants engaging *more* with the like button in the Delayed condition. We found a significant effect of immediacy on user engagement as measured by the number times people clicked the like button in our online study ($p < 10^{-16}$). As the solid bars in Figure 5 show, when engagement is measured only by likes, people appeared to be more engaged when updates to recommendations are delayed. A post-hoc analysis confirmed this – the two conditions with delayed updates had significantly more like actions than all others. Compared to the Instantaneous or Never conditions with around 4.8 like actions on average, there was 55% or 98% increase in the number of actions in Delayed conditions (with means 7.5 and 8.5). There were no significant differences between any of the Preview, Instantaneous or Never conditions with respect to the number of like actions.

Although engagement through actions such as liking is often considered an indication of user endorsement of recommender systems [13], participant comments from our lab study suggested the opposite:

- *"I was doing it because I was trying to get my recommendations to change."* (P7)
- *"I felt like more obligated to click ... It needed more clicks to do something."* (P35)
- *"I am probably engaging more because it is the only way to get the suggested stories up."* (P40)

On deeper examination, we find evidence that focusing solely on likes as user engagement can paint a misleading picture. We found taking likes (explicit feedback), hovers over the like button (an indication of consideration) and readlisting actions (a type of conversion in this task) to be a more robust indication of overall engagement (see Figure 5). The overall quantity of these actions matters because those actions would be typically taken as labeled examples in training the recommender system and is directly linked to predictive performance of the recommender system [4]. As the cross-hatched bars show, the overall engagement is comparable across all conditions. Connecting this with the increased engagement with the like button implies that the extra likes in the delayed condition did not lead to an overall increase in available information about a user's preferences as the information was likely captured by a readlist or view action as well. In fact, as can be seen from the lab participants' comments above, the likes were not necessarily an indicator of interest but rather a way to get the recommendations to change. Some participants in our lab study even commented about not seeing value in engaging with the like button in the Delayed condition:

- *"Too much work, getting too little out of it."* (P8)
- *"When there is no action right away, it feels like not worth doing sometimes."* (P9)

Visible changes in the Diff condition were strongly preferred for a perceived increase in responsiveness and transparency, but destructive changes should be avoided. In comparing the two visibility conditions, 9 out of 11 of our lab study participants reported preferring the Diff condition which explicitly and visibly highlighted changes to the recommendations. This is compared to the Refresh condition which simply refreshed the recommendations after an update. The main justifications participants gave for this preference was that the Diff condition made clear that something was happening as a result of their actions (i.e., it increased the perceived responsiveness of the system) and, moreover, made clear what was happening:

- *"This made it really clear what was new and what was not."* (P8)
- *"I could see action happening ... It felt like something was getting done. The system was giving me feedback for my actions that I was taking ... At a high level, the system was reacting to me better"* (P9)
- *"I want the algorithm to signal the action that it has taken – so that there is almost more of a conversation between me and it"* (P31)

In contrast, participants commented that they were unsure if and how the system was updating in the Refresh condition or getting better over time:

- *"[In the Refresh condition], I wasn't sure whether anything refreshed or changed."* (P33)
- *"When they change I want to see the highlight ... [with the Refresh condition] I felt like things were moving around but I couldn't keep track."* (P2)
- *"[In the Refresh condition], I was trying to trigger a change, but it was hard to tell whether it was changing."* (P7)

Preferences for the Diff condition were less discernible from participant behaviors alone in either the lab or the online studies. Analysis of behaviors in our online study showed a slight increase in engagement with recommendations in the Diff condition compared to the Refresh condition (see Figure 4), suggesting it drew more attention from participants, but this increase was subtle with no significant differences observed.

While most participants preferred the Diff condition in our lab study, several distinguished between the constructive and destructive changes (highlighting newly added items in yellow or greying out items that will be removed, respectively) with most stating that they would prefer only constructive changes:

- *"I am immediately looking at what gets greyed out… would prefer if the greyed out items stayed."* (P40)
- *"But the way that it was presented was a bit to harsh and it was difficult to read the greyed out ones."* (P16)
- *"With the greyed out ones, I was wasting time on reading what was going away … It made me panic a bit."* (P40)

8 DISCUSSION

We will discuss the broader findings of our study as well as their implications for improving current recommender systems. As the qualitative study revealed, there was a clear preference for Preview in terms of when a system should respond to a user action, and Diff regarding the way it presents changes to users. Preview was preferred over the more traditional UX in most recommender systems where people cannot anticipate the impact of their actions on the recommendation system – for example Instantaneous and Delayed. Preview was mentioned to improve the sense of control of the system's behavior and reduce decision anxiety. In addition to improved user experience, offering Preview to people increased selectivity in the like actions, creating an opportunity to use more fine-grained feedback for the ML. For example, it might be beneficial to distinguish between items where the like button was clicked and items where the like button was only hovered over. Moreover, given the occurrence of peeking behavior, one could use previewed but not persisted items as negative training instances. Implementing Preview in real systems might be computationally more challenging compared to the other conditions, but is more than feasible given current client-side computing power [50].

Our qualitative study also showed that people strongly preferred to have visibly highlighted changes, improving perceived responsiveness and transparency. This emphasizes the importance of supporting people in evaluating outcomes during decision making [27, 38]. Our quantitative study showed subtle effects of highlighting in Diff on attention – there was a small but not statistically significant increase in the probability to engage with new recommendations via hover or clicks when compared to Preview Refresh. Preview Diff did have a statistically significant increase when compared to *any* of the Refresh interfaces – which represent the baselines in almost all applications. Given that implicit behavior signals are often an approximation of gaze and attention, it would be interesting to repeat the study with eye-tracking devices for a more in-depth study of attention. Even though the overall feedback around Diff was positive, some types of highlighted changes can be distracting, and in particular, destructive changes were reported to be disconcerting. We may be able to improve even further by only highlighting additions to the recommendations and supporting an infinite scroll where deprioritized recommendations based on feedback actions gracefully disappear below the fold.

Our studies also showed that number of likes as the sole measure of engagement is a noisy signal and requires careful interpretation as it may indicate the opposite of what is intended. For example, even though engagement with the like button increased significantly in the Delayed conditions, this was not indicative of overall user preference. A perspective of engagement that considers hovers on likes and readlisting in addition to likes was more robust across conditions. In conditions where changes could not be previewed, users clicked "like" both to elicit a change from the system and to express preference. The Never condition gives an even more drastic example – people still interacted with the like button, even though it never reacted to any of the actions. In contrast, in Preview users clicked "like" more frequently because they felt expressing their preferences changed the recommendations in a way they wanted. This provides a stronger link from user feedback to recommendation behavior, and users in this condition demonstrated satisfaction by engaging with recommendation much more frequently after an update in the Preview condition. In summary, we found that the motivations people have for liking things depends in part on how the system and UX responds. Failure to consider this can confound the interpretation of implicit behavior signals as evidence of what people like. This highlights the importance of a multi-perspective approach to the evaluation of recommender systems.

9 LIMITATIONS AND FUTURE WORK

One limitation of the studies in this paper is that they were using a desktop environment with a mouse. This enabled us to use hovering as an interaction technique in the Preview condition. However, we believe that similar interaction experiences can also be implemented on touch-only devices, whether it is through the use of tapping, force-touch or more sophisticated techniques [49].

The other limitations concern the length of the study – both regarding the task as well as the overall study session due to constraints from the lab and crowd-sourced environments. Time constraints may impact satisfaction and strategies employed [8]. Thus, in the future, it would be interesting to repeat this study in an unconstrained, longitudinal setting in order to study long-term outcomes such as repeated usage and trust.

Future work can expand on our findings in multiple ways. We focused on immediacy and visibility as two key dimensions of supporting user control in recommender systems, but there are other important aspects, for example explanations [44].Moreover, with the results stressing the importance of design guidelines for AI systems around conveying consequences of user actions, it would be interesting to also explore other design factors, such as incremental and more frequent vs. larger and less frequent changes to the AI model [2].

10 CONCLUSIONS

In this paper, we studied two factors that aim at helping people understand how their feedback actions will impact their personalized recommendations – the immediacy and visibility of the system update. Our results show that regarding immediacy, having a previewing mechanism was favored by people over other more delayed updates common in most recommender systems. Having the ability to preview updates, increased feelings of control and reduced decision anxiety. Moreover, we found that previewing increased selectivity in what people like, thereby providing the potential to use a more fine-grained signal for training and evaluation. In terms of visibility, the overall majority of participants preferred having changes highlighted. This was mostly due to a perceived increase in transparency and responsiveness.

We hope that this paper will help inform future research in personalization and recommendation. In particular, in this paper we studied only two facets of improving user control through a better understanding of the consequences of their feedback in interactive human-in-the-loop systems and hope that this will inspire more research in this largely unexplored area [6, 10].

This work was funded in part by NSF awards IIS-1513692 and IIS-1901168.

REFERENCES

[1] Xavier Amatriain and Justin Basilico. 2015. Recommender systems in industry: A Netflix case study. In *Recommender Systems Handbook*. Springer, 385–419.

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-ai interaction. In *CHI*. 1–13.

[3] Fedor Bakalov, Marie-Jean Meurs, Birgitta König-Ries, Bahar Sateli, René Witte, Greg Butler, and Adrian Tsang. 2013. An approach to controlling user models and personalization effects in recommender systems. In *IUI*. 49–56.

[4] JesúS Bobadilla, Fernando Ortega, Antonio Hernando, and JesúS Bernal. 2012. A collaborative filtering approach to mitigate the new user cold start problem. *Knowledge-based systems* 26 (2012), 225–238.

[5] Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. 2012. TasteWeights: a visual interactive hybrid recommender system. In *RecSys*. 35–42.

[6] Eli T. Brown, Remco Chang, and Alex Endert. 2016. Human-Machine-Learner Interaction: The Best of Both Worlds. In *CHI Workshop on Human Centered Machine Learning*.

[7] Dan Cosley, Shyong K Lam, Istvan Albert, Joseph A Konstan, and John Riedl. 2003. Is seeing believing?: how recommender system interfaces affect users' opinions. In *CHI*. 585–592.

[8] Anita Crescenzi, Diane Kelly, and Leif Azzopardi. 2016. Impacts of time constraints and system delays on user experience. In *SIGIR*. 141–150.

[9] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41, 6 (1990), 391–407.

[10] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. In *CHI*. 278–288.

[11] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I like it, then I hide it: Folk theories of social feeds. In *CHI*. 2371–2382.

[12] James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. 2008. CueFlik: interactive concept learning in image search. In *CHI*. 29–38.

[13] Florent Garcin, Boi Faltings, Olivier Donatsch, Ayar Alazzawi, Christophe Bruttin, and Amr Huber. 2014. Offline and online evaluation of news recommender systems at swissinfo. ch. In *RecSys*. 169–176.

[14] Donald A Hantula, Diane DiClemente Brockman, and Carter L Smith. 2008. Online shopping as foraging: The effects of increasing delays on purchasing and patch residence. *IEEE Transactions on Professional Communication* 51, 2 (2008), 147–154.

[15] Chen He, Denis Parra, and Katrien Verbert. 2016. Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications* 56 (2016), 9–27.

[16] Dietmar Jannach, Sidra Naveed, and Michael Jugovac. 2016. User control in recommender systems: Overview and interaction challenges. In *International Conference on Electronic Commerce and Web Technologies*. 21–33.

[17] Yucheng Jin, Bruno Cardoso, and Katrien Verbert. 2017. How do different levels of user control affect cognitive load and acceptance of recommendations?. In *RecSys Workshop on Interfaces and Human Decision Making for Recommender Systems*. 35–42.

[18] Michael Jugovac and Dietmar Jannach. 2017. Interacting with recommenders—overview and research directions. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7, 3 (2017), 10.

[19] Yvonne Kammerer, Rowan Nairn, Peter Pirolli, and Ed H Chi. 2009. Signpost from the masses: learning effects in an exploratory social tag search browser. In *CHI*. 625–634.

[20] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *CHI*. 453–456.

[21] Bart P Knijnenburg, Svetlin Bostandjiev, John O'Donovan, and Alfred Kobsa. 2012. Inspectability and control in social recommenders. In *RecSys*. 43–50.

[22] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *IUI*. 126–137.

[23] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *WWW*. 661–670.

[24] Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. 2011. Content-based recommender systems: State of the art and trends. In *Recommender Systems Handbook*. Springer, 73–105.

[25] Sean M McNee, Shyong K Lam, Joseph A Konstan, and John Riedl. 2003. Interfaces for eliciting new user preferences in recommender systems. In *UMAP*. 178–187.

[26] Jakob Nielsen. 1994. *Usability engineering*. Elsevier.

[27] Jakob Nielsen. 1995. 10 usability heuristics for user interface design. *Nielsen Norman Group* 1, 1 (1995).

[28] Jakob Nielsen. 1999. User interface directions for the Web. *Commun. ACM* 42, 1 (1999), 65–72.

[29] Donald A Norman. 1988. *The psychology of everyday things*. Basic books.

[30] John O'Donovan, Barry Smyth, Brynjar Gretarsson, Svetlin Bostandjiev, and Tobias Höllerer. 2008. PeerChooser: visual interactive recommendation. In *CHI*. 1085–1088.

[31] Denis Parra and Peter Brusilovsky. 2015. User-controllable personalization: A case study with SetFusion. *International Journal of Human-Computer Studies* 78 (2015), 43–67.

[32] Pearl Pu and Li Chen. 2006. Trust building with explanation interfaces. In *IUI*. 93–100.

[33] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *RecSys*. 157–164.

[34] Pernilla Qvarfordt, Gene Golovchinsky, Tony Dunnigan, and Elena Agapie. 2013. Looking ahead: query preview in exploratory search. In *SIGIR*. 243–252.

[35] Joseph John Rocchio. 1971. Relevance feedback in information retrieval. *The SMART retrieval system: experiments in automatic document processing* (1971), 313–323.

[36] J Ben Schafer, Joseph A Konstan, and John Riedl. 2002. Meta-recommendation systems: user-controlled integration of diverse recommendations. In *CIKM*. 43–51.

[37] James Schaffer, Tobias Hollerer, and John O'Donovan. 2015. Hypothetical recommendation: A study of interactive profile manipulation behavior for recommender systems. In *FLAIRS*.

[38] Donald A Schön. 2017. *The reflective practitioner: How professionals think in action*. Routledge.

[39] Ben Shneiderman. 1984. Response time and display rate in human performance with computers. *Comput. Surveys* 16, 3 (1984), 265–285.

[40] Ben Shneiderman and Pattie Maes. 1997. Direct manipulation vs. interface agents. *Interactions* 4, 6 (1997), 42–61.

[41] Jacob Solomon. 2014. Customization bias in decision support systems. In *CHI*. 3065–3074.

[42] E Isaac Sparling and Shilad Sen. 2011. Rating: how difficult is it?. In *RecSys*. 149–156.

[43] Michael Terry and Elizabeth D Mynatt. 2002. Side views: persistent, on-demand previews for open-ended tasks. In *UIST*. 71–80.

[44] Nava Tintarev and Judith Masthoff. 2007. A survey of explanations in recommender systems. In *International Conference on Data Engineering*. 801–810.

[45] Kristen Vaccaro, Dylan Huang, Motahhare Eslami, Christian Sandvig, Kevin Hamilton, and Karrie Karahalios. 2018. The Illusion of Control: Placebo Effects of Control Settings. In *CHI*. 16.

[46] Flavian Vasile, Elena Smirnova, and Alexis Conneau. 2016. Meta-Prod2Vec: Product embeddings using side-information for recommendation. In *RecSys*. 225–232.

[47] Katrien Verbert, Denis Parra, Peter Brusilovsky, and Erik Duval. 2013. Visualizing recommendations to support exploration, transparency and controllability. In *IUI*. 351–362.

[48] Wesley Waldner and Julita Vassileva. 2014. Emphasize, don't filter!: displaying recommendations in Twitter timelines. In *RecSys*. 313–316.

[49] Feng Wang and Xiangshi Ren. 2009. Empirical evaluation for finger input properties in multi-touch interaction. In *CHI*. 1063–1072.

[50] Chao-Yuan Wu, Christopher V Alvino, Alexander J Smola, and Justin Basilico. 2016. Using navigation to improve recommendations in real-time. In *RecSys*. 341–348.